

Normalized Ordinal Distance; A Performance Metric for Ordinal, Probabilistic-ordinal or Partial-ordinal Classification Problems

Mohammad Hasan Bahari, Hugo Van hamme

Center for processing speech and images, KU Leuven, Belgium

{mohamadhasan.bahari, hugo.vanhamme}@esat.kuleuven.be

Abstract

In many forensic scenarios, we deal with problems of ordinal nature, where there is intrinsic ordering between the categories. For example, in human age group recognition from speech or images, the categories can be child, young, middle-aged and senior. In detecting the level of intoxication, the categories can be low, medium and high. In this paper, a novel application-independent performance metric for ordinal, probabilistic-ordinal and partial-ordinal classification problems is introduced. Conventional performance metrics for ordinal classification problems, such as mean absolute error of consecutive integer labels and ranked probability score, are difficult to interpret and can be misleading. In this paper, first, the ordinal distance between two arbitrary vectors in Euclidean space is introduced. Then, a new performance metric, namely normalized ordinal distance, is proposed based on the introduced ordinal distance. This performance metric is conceptually simple, computationally inexpensive and application-independent. The advantages of the proposed method over the conventional approaches and its different characteristics are shown using several numerical examples.

1. Introduction

A large number of real world classification problems are ordinal, where there is intrinsic ordering between the categories. For example, in quality prediction systems, the task is to categorize the quality of a product into bad, good and excellent [1]. In human age group recognition from speech or images, the categories can be child, young, middle-aged and senior [2, 3]. In the classification of the therapeutic success, the classes are good recovery, moderate disability, severe disability, and fatal outcome [4]. In all ordinal classification problems (C_O), the class labels are ordinal numbers, i.e. there is intrinsic ordering between the categories.

Probabilistic-Ordinal and Partial-Ordinal Classification problems, labeled as C_O^{Pr} and C_O^{Pa} respectively, are well-known generalizations of the C_O . In C_O^{Pr} , for a test datapoint, the classifier calculates the probability of belonging to each category. In C_O^{Pa} , instead of the crisp class labels each datapoint has a degree of membership

to every class [5]. These types of problems, explained in sections 2.2 and 2.3 in detail, can be found in many domains, such as natural language processing, social network analysis, bioinformatics and agriculture [5].

Scientists have proposed different methods to solve C_O , C_O^{Pr} and C_O^{Pa} [5, 6, 7, 8, 9, 10]. For example, McCullagh introduced an ordinal classifier, namely the proportional odds model (POM), based on logistic regression [6]. In [7], C_O is addressed using a generalization of support vector machines (SVM) namely support vector ordinal regression (SVOR). A neural network approach for the C_O is suggested in [8]. [9] suggested Gaussian processes for C_O . In [5], kernel-based proportional odds models is introduced to solve the C_O^{Pa} .

To measure the performance of these classifiers, different approaches have been suggested. For example, mean zero-one error (E_{mzo}) and mean absolute error of consecutive integer labels (E_{ma}^{cil}) are widely applied to measure the performance of the classifiers in C_O [7, 8, 9, 10]. However, non of these methods are applicable to C_O^{Pr} and C_O^{Pa} . Percentage of correctly fuzzy classified instances (P_{cfci}) and Average Deviation (E_{ad}) have been suggested to measure the classifier performance in C_O^{Pr} and C_O^{Pa} [5, 11, 12, 13]. The main drawback of P_{cfci} is that it does not consider the order of categories [11, 12]. The E_{ad} suggests a simple idea to solve this problem [12, 13]. Although the E_{ad} is attractive from several aspects, the interpretation of its results is difficult, because the range of its output depends on the application. The same difficulty is observed in E_{ma}^{cil} . Application dependency makes the interpretation of E_{ma}^{cil} and E_{ad} very challenging. The average of ranked probability scores (E_{tps}), is also applied as a performance metric in C_O^{Pr} and C_O^{Pa} [14, 15]. In this method, the order of categories is important and the range of the output is fixed between 0 and 1. This method can be applied to C_O , C_O^{Pr} and C_O^{Pa} . However, analysis reveals that E_{tps} over estimates the performance of classifiers in many situations. This issue, which leads to a erroneous interpretation of classifier performance, is illustrated by some numerical examples in section 5.

In this paper, we investigate different characteristics of these performance metrics and finally a new application-independent performance metric, namely Normalized Or-

dinal Distance (E_{nod}^p), is introduced. The Matlab code of the suggested approach, which can be applied to all three types of considered problems C_O , C_O^{Pr} and C_O^{Pa} , can be downloaded from our website¹.

This paper is organized as follows. In section 2, the mathematical formulations of C_O , C_O^{Pr} and C_O^{Pa} are presented. In section 3, five different conventional performance metrics are explained. The proposed performance metric is elaborated in section 4. In section 5, the effectiveness of the proposed approach is illustrated using some numerical examples. The paper ends with a conclusion in section 6.

2. Problem Formulation

In this section, the ordinal, probabilistic-ordinal and partial-ordinal problems are formulated.

2.1. Ordinal Classification

Assume that we are given a training data set $S^{\text{tr}} = \{(X_1, Y_1), \dots, (X_n, Y_n), \dots, (X_N, Y_N)\}$, where $X_n = [x_{n,1}, \dots, x_{n,i}, \dots, x_{n,L}]$ denotes a vector of observed characteristics of the data item and $Y_n = [y_{n,1}, \dots, y_{n,d}, \dots, y_{n,D}]$ denotes a label vector. The label vector is defined as follows if X_n belongs to class C_d :

$$y_{n,j} = \delta_{j,d}. \quad (1)$$

where δ denotes the Kronecker delta. In ordinal problems, there is an intrinsic ordering between the classes, which is denoted as $C_1 \prec \dots \prec C_d \prec \dots \prec C_D$ like low, medium and high [5]. The goal is to approximate a classifier function (G), such that for the m^{th} unseen observation X_m^{tst} , $\hat{Y}_m = G(X_m^{\text{tst}})$ is as close as possible to the true label. For a crisp classifier \hat{Y}_m is defined as follows if the d^{th} class is chosen for X_m^{tst}

$$\hat{y}_{m,j} = \delta_{j,d}. \quad (2)$$

2.2. Probabilistic-Ordinal classification

Probabilistic-ordinal classification problem (C_O^{Pr}) is a generalization of the C_O , where each element of the classifier output vector (\hat{Y}) represents the probability of belonging to the corresponding category. In this type of classification, Y_n is defined by relation (1). However, \hat{Y}_m is defined as follows

$$\hat{Y}_m = \left\{ \begin{array}{l} [\hat{y}_{m,1}, \dots, \hat{y}_{m,d}, \dots, \hat{y}_{m,D}] \in \mathbb{R}^D \\ \hat{y}_{m,d} \geq 0; \sum_{d=1}^D \hat{y}_{m,d} = 1 \end{array} \right\}, \quad (3)$$

where \mathbb{R} denotes the set of real numbers.

2.3. Partial-Ordinal Classification

Partial-ordinal classification problem (C_O^{Pa}) is another generalization of C_O [5]. In ordinal problems, each data object

is limited to belong to a single category, i.e. out of all D elements of Y_n , only one is nonzero. However, this is too conservative in the case of non-crisp or fuzzy classes. This limitation is relaxed in C_O^{Pa} by rephrasing Y_n as follows

$$Y_n = \left\{ \begin{array}{l} [y_{n,1}, \dots, y_{n,d}, \dots, y_{n,D}] \in \mathbb{R}^D \\ y_{n,d} \geq 0; \sum_{d=1}^D y_{n,d} = 1 \end{array} \right\}. \quad (4)$$

Therefore, each datapoint has a degree of membership to all classes. Like in ordinal problems, the final goal is to approximate a classifier function (G), such that for an unseen observation X^{tst} , $\hat{Y}_m = G(X_m^{\text{tst}})$ is as close as possible to the true label. In this type of classification \hat{Y}_m is also defined by relation 3.

3. Conventional Performance Metrics

In this section, five widely-used conventional metrics, namely E_{mzo} , $E_{\text{ma}}^{\text{cil}}$, P_{fci} , E_{ad} and E_{rps} are introduced [5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17].

3.1. Mean Zero-One Error (E_{mzo})

Performance metric E_{mzo} is the fraction of incorrect predictions, which is calculated as follows [7, 8, 9, 10]

$$E_{\text{mzo}} = \frac{1}{M} \sum_{m=1}^M (1 - \delta_{\hat{y}_m, y_m}), \quad (5)$$

where M is the total number of test set datapoints, \hat{y}_m is the predicted label of the m^{th} test set datapoint and y_m is the true label of the m^{th} test set datapoint. The main advantage of E_{mzo} is its simplicity. However, it does not consider the order of the categories. Furthermore, it is not applicable to measure the performance in C_O^{Pr} or C_O^{Pa} .

3.2. Mean Absolute Error of Consecutive Integer Labels ($E_{\text{ma}}^{\text{cil}}$)

To calculate the $E_{\text{ma}}^{\text{cil}}$, first, both true labels and predicted labels of the test set datapoints are transformed into consecutive integers so that if the d^{th} column of the label vector is 1 then the transformed label is equal to d [7, 8, 9, 10]. After label transformation the $E_{\text{ma}}^{\text{cil}}$ is calculated as follows

$$E_{\text{ma}}^{\text{cil}} = \frac{1}{M} \sum_{m=1}^M |\hat{U}_m - U_m|, \quad (6)$$

where \hat{U}_m is the transformed predicted label of the m^{th} test set datapoint and U_m is the transformed true label of the m^{th} test set datapoint. The $E_{\text{ma}}^{\text{cil}}$ enjoys the advantage of considering the order of categories into account. However, it cannot be applied to evaluate the classifiers in C_O^{Pr} or C_O^{Pa} . Moreover, the range of its output is application-dependent. Therefore, the interpretation of this metric is challenging. This is shown in section 5 using some numerical examples.

¹<http://www.esat.kuleuven.be/psi/spraak/downloads/>

3.3. Percentage of Correctly Fuzzy Classified Instances (P_{cfci})

Performance metric P_{cfci} has been applied to measure the performance of probabilistic or fuzzy classifiers [11, 12]. It is calculated as follows

$$P_{\text{cfci}} = \frac{100}{M} \sum_{m=1}^M \left(1 - \frac{1}{2} \sum_{d=1}^D |\hat{y}_{m,d} - y_{m,d}|\right). \quad (7)$$

As it can be inferred from the above relation, the order of the categories is not considered in P_{cfci} .

3.4. Average Deviation (E_{ad})

Performance metric E_{ad} was originally introduced by Van Broekhoven [12] to evaluate the classifiers in fuzzy ordered classification problems. It was also applied in different applications with other names [5, 13]. The E_{ad} is calculated as follows

$$E_{\text{ad}} = \frac{1}{M} \sum_{m=1}^M \left\{ \sum_{d=1}^{D-1} \left| \sum_{i=1}^d \hat{y}_{m,i} - \sum_{i=1}^d y_{m,i} \right| \right\}. \quad (8)$$

It can be interpreted from the above relation that the order of categories is important in E_{ad} . It is also useful for classifier evaluation in C_{O}^{Pr} or C_{O}^{Pa} . However, similar to $E_{\text{ma}}^{\text{cil}}$, the range of E_{ad} is application-dependent and hence difficult to interpret.

3.5. Average Ranked Probability Scores (E_{rps})

The ranked probability score was originally introduced to score the output of probabilistic classifiers [14, 15]. It is defined as follows

$$RPS_Y(\hat{Y}) = \frac{1}{D-1} \left\{ \sum_{d=1}^{D-1} \left(\sum_{i=1}^d \hat{y}_i - \sum_{i=1}^d y_i \right)^2 \right\}. \quad (9)$$

This scoring rule can be easily extended to measure the performance of classifiers in C_{O} , C_{O}^{Pr} and C_{O}^{Pa} using the following relation.

$$E_{\text{rps}} = \frac{1}{M(D-1)} \sum_{m=1}^M \sum_{d=1}^{D-1} \left(\sum_{i=1}^d \hat{y}_{m,i} - \sum_{i=1}^d y_{m,i} \right)^2 \quad (10)$$

As it can be interpreted from the above relation, the order and the number of categories are important in E_{rps} . It is assumed that the maximum of the nominator of E_{rps} is $M(D-1)$. Therefore, to fix the range of E_{rps} between 0 and 1 the nominator is divided to its maximum possible value $M(D-1)$. However, this assumption is very conservative so that in many practical cases the maximum of the nominator of E_{rps} is less than $M(D-1)$. Consequently, this assumption may lead to an erroneous interpretation of the classifier performance. Numerical examples of Section 5 reveal this issue clearly.

4. Proposed Performance Metric

In this section, first, Ordinal Distance (OD) of two vectors in Euclidean space is introduced. Then, a new performance metric, namely normalized ordinal distance (E_{nod}^p), is developed based on the ordinal distance.

4.1. Ordinal Distance (OD)

In this section, the definition of a distance function is recaptured. Then, the Minkowski distance is described and finally, the ordinal distance is introduced as an extension of the Minkowski distance.

4.1.1. Distance

By definition, a distance function of two points $A = [a_1, \dots, a_d, \dots, a_D]$ and $B = [b_1, \dots, b_d, \dots, b_D]$ is a function $D: \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$, which satisfies the following three conditions [18]:

1. $D(A, B) \geq 0$ and $D(A, B) = 0 \Leftrightarrow A = B$
2. $D(A, B) = D(B, A)$
3. $D(A, C) \leq D(A, B) + D(B, C)$

A variety of distance functions have been introduced by scientists for different applications such as Minkowski distance, Mahalanobis distance, Chebyshev distance and Hamming distance [18].

4.1.2. The Minkowski distance of order p

The Minkowski distance of order p or p -norm is a distance function, which satisfies all conditions of a distance function.

$$\|A - B\|_p = \left(\sum_{d=1}^D |a_d - b_d|^p \right)^{1/p} \quad (11)$$

where p is a real number not less than 1. As in can be interpreted from relation (11), in p -norm, the order of the elements of two points A and B , is not important.

4.1.3. The Ordinal distance of order p

The notion of ordinal distance is previously used to measure the differences of two strings [19] or two histograms [20]. In this paper, an ordinal distance of two vectors in Euclidean space is introduced. The Ordinal Distance of order p between two points A and B is defined as follows

$$\begin{aligned} \|A - B\|_p^{\text{OD}} &= \left(\sum_{d=1}^D |\bar{a}_d - \bar{b}_d|^p \right)^{1/p} \\ \bar{a}_d &= \sum_{i=1}^d a_i \\ \bar{b}_d &= \sum_{i=1}^d b_i, \end{aligned} \quad (12)$$

where p is a real number not less than 1. Since (12) is a Minkowski distance between $\bar{A} = [\bar{a}_1 \dots \bar{a}_d \dots \bar{a}_D]$ and $\bar{B} = [\bar{b}_1 \dots \bar{b}_d \dots \bar{b}_D]$, it follows that the ordinal distance of order p satisfies the conditions of section 4.1.1.

4.2. Normalized Ordinal Distance (E_{nod}^p)

In this section, a new performance metric, namely normalized ordinal distance (E_{nod}^p), is introduced to measure the performance classifiers in C_O , C_O^{Pr} and C_O^{Pa} .

$$E_{\text{nod}}^p = \frac{\sum_{m=1}^M \|Y_m - \hat{Y}_m\|_p^{\text{OD}}}{\sum_{m=1}^M \psi_{Y_m}^p}, \quad (13)$$

where $\psi_{Y_m}^p$ is the upper bound of $\|Y - \hat{Y}\|_p^{\text{OD}}$ for any possible \hat{Y} in its defined range. ψ_Y is defined as follows

$$\psi_Y^p \triangleq \max_T \|Y - T\|_p^{\text{OD}}, \quad (14)$$

where $T = \{t_1, \dots, t_d, \dots, t_D\}$ is an arbitrary vector with the same specifications of \hat{Y} mentioned in relation (2), i.e., T lies on a simplex. ψ_Y^p can be calculated using theorem 1.

Theorem 1:

The upper bound of $\|Y - \hat{Y}\|_p^{\text{OD}}$ for any possible \hat{Y} can be obtained as follows

$$\psi_Y^p = \max(\|Y - L_1\|_p^{\text{OD}}, \dots, \|Y - L_d\|_p^{\text{OD}}, \dots, \|Y - L_D\|_p^{\text{OD}}) \quad (15)$$

or equivalently

$$\psi_Y^p = \max(\|Y - L_1\|_p^{\text{OD}}, \|Y - L_D\|_p^{\text{OD}}), \quad (16)$$

where L_d is a vector of size Y . The d^{th} element of L_d is equal to 1 and the rest of elements are zero. As it can be interpreted from relations (15) and (16), although the latter one is more restrictive, it provides an easier way to calculate ψ_Y^p .

Proof:

We first prove the relation (15), which help us to show the correctness of relation (16).

Proof of relation (15):

By definition

$$\|Y - T\|_p^{\text{OD}} = \|\Lambda(Y - T)\|_p, \quad (17)$$

where Λ is a lower triangular matrix of size $D \times D$ with all diagonal and lower diagonal elements equal to 1. Since $\|(Y - T)\|_p$ is a convex function of T and a convex function remains convex under an affine transformation, $\|\Lambda(Y - T)\|_p$ is also convex.

On the other hand, a convex function on a compact convex set attains its maximum at an extreme point of the set [21]. In this problem $T \in \{[t_1, \dots, t_d, \dots, t_D] \in \mathbb{R}^D | t_d \geq 0; \sum_{d=1}^D t_d = 1\}$, i.e., T lies on a simplex. The extreme points of this compact convex set are L_d with

$d \in \{1, \dots, D\}$.

Therefore,

$$\max_T \|\Lambda(Y - T)\|_p = \max(\|\Lambda(Y - L_1)\|_p, \dots, \|\Lambda(Y - L_d)\|_p, \dots, \|\Lambda(Y - L_D)\|_p) \quad (18)$$

Consequently,

$$\max_T \|Y - T\|_p^{\text{OD}} = \max(\|Y - L_1\|_p^{\text{OD}}, \dots, \|Y - L_d\|_p^{\text{OD}}, \dots, \|Y - L_D\|_p^{\text{OD}}) \quad (19)$$

Proof of relation (16):

Relation (16) is now shown by contradiction. Suppose relation (15) is not equivalent with relation (16), then there must be a $k \in \{2, \dots, D-1\}$ such that both Relations 20 nad 21 hold.

$$\|Y - L_k\|_p^{\text{OD}} > \|Y - L_1\|_p^{\text{OD}} \quad (20)$$

$$\|Y - L_k\|_p^{\text{OD}} > \|Y - L_D\|_p^{\text{OD}} \quad (21)$$

Expansion of relation (20) and (21) is

$$\sum_{d=1}^{k-1} \left(\sum_{i=1}^d y_i \right)^p + \sum_{d=k}^{D-1} \left(1 - \sum_{i=1}^d y_i \right)^p > \sum_{d=1}^{D-1} \left(1 - \sum_{i=1}^d y_i \right)^p \quad (22)$$

$$\sum_{d=1}^{k-1} \left(\sum_{i=1}^d y_i \right)^p + \sum_{d=k}^{D-1} \left(1 - \sum_{i=1}^d y_i \right)^p > \sum_{d=1}^{D-1} \left(\sum_{i=1}^d y_i \right)^p. \quad (23)$$

After some manipulations (22) and (23) lead to

$$\sum_{d=1}^{k-1} \left[\left(\sum_{i=1}^d y_i \right)^p - \left(1 - \sum_{i=1}^d y_i \right)^p \right] > 0 \quad (24)$$

$$\sum_{d=k}^{D-1} \left[\left(1 - \sum_{i=1}^d y_i \right)^p - \left(\sum_{i=1}^d y_i \right)^p \right] > 0. \quad (25)$$

If relation (24) holds, $(\sum_{i=1}^d y_i) > (1 - \sum_{i=1}^d y_i)$ hence $(\sum_{i=1}^d y_i) > 0.5$ for at least one d between 1 and $k-1$. Likewise, from (25), $(\sum_{i=1}^d y_i) < 0.5$ for at least one d between k and $D-1$. This is impossible, since $\sum_{i=1}^d y_i$ is an increasing function of d and hence (16) holds.

4.2.1. Relation to E_{tps}

There is a close relationship between E_{tps} and E_{nod}^p specially for $p = 2$. In both E_{tps} and E_{nod}^p , denominators are assumed to be the upper bound of the nominator and are used to keep the range of performance metric between 0 and 1. In E_{tps} , it is assumed that the upper bound of the nominator is $M(D-1)$ [15, 22]. However, this is a conservative bound in many situations. This is illustrated by some numerical examples in section 5. We will also show this conservative assumption can result in a misleading or erroneous interpretation of the classifiers performance. In E_{nod}^p , this upper bound is explicitly defined by relation (14) and calculated by relation (16).

Table 1: The performance of two classifiers measured by E_{mzo} , E_{ad} , E_{ma}^{cil} , P_{cfci} , E_{rps} , E_{nod}^1 , E_{nod}^2 , and E_{nod}^∞ in example 1.

Performance Metric	Problem 1	Problem 2
E_{mzo}	0.05	0.05
E_{ad}	0.05	0.1
E_{ma}^{cil}	0.05	0.1
P_{cfci}	97.5	97.5
E_{rps}	0.025	0.05
E_{nod}^1	0.0286	0.0571
E_{nod}^2	0.0381	0.0540
E_{nod}^∞	0.05	0.05

5. Results and Discussion

In this section, the behaviors of E_{nod}^p and five conventional performance metrics, namely E_{mzo} , P_{cfci} , E_{ad} , E_{rps} , and E_{ma}^{cil} are analyzed using a number of numerical examples. In example 1, it is shown that P_{cfci} and E_{mzo} are not suitable for measuring the performance of ordinal classifiers, because these methods do not consider the order of categories. Examples 2 illustrates that the interpretation of E_{ad} and E_{ma}^{cil} depends on the number of categories, hence these metrics are application-dependent. In example 3, the deficiency of E_{rps} is demonstrated. Examples 4 shows the advantages of E_{nod}^p over P_{cfci} , E_{rps} and E_{ad} in measuring the performance of the classifiers in C_O^{Pa} , where other conventional approaches are not applicable.

5.1. Example 1

In this example, the advantage of E_{nod}^p over P_{cfci} and E_{mzo} in measuring the performance of the classifiers in a typical C_O is shown. For an ordinal three-class classification problem, classifier 1 and classifier 2 result in confusion matrix 1, labeled as CM_1 and CM_2 respectively. In these matrices each column represents the instances in a predicted class and each row shows the instances in an actual class.

$$CM_1 = \begin{bmatrix} 4 & 1 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 10 \end{bmatrix} \quad CM_2 = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 5 & 0 \\ 0 & 0 & 10 \end{bmatrix} \quad (26)$$

Table 1 shows the performance of two classifiers measured by E_{mzo} , E_{ad} , E_{ma}^{cil} , P_{cfci} , E_{rps} , E_{nod}^1 , E_{nod}^2 , and E_{nod}^∞ . As it can be interpreted from this table, E_{mzo} , P_{cfci} and E_{nod}^∞ fail to reflect the degradation of performance from the classifier 1 to the classifier 2. However, E_{nod}^1 , E_{nod}^2 , E_{ad} , E_{rps} and E_{ma}^{cil} perfectly show that classifier 1 outperforms classifier 2.

5.2. Example 2

In this example, we show that the number of categories in the classification problem influences the interpretation of E_{ad} and E_{ma}^{cil} . Consider the following two ordinal classification problems.

Problem 1: For a test datapoint, the true label and the estimated label are $Y_1 = [1 \ 0]$ and $\hat{Y}_1 = [0 \ 1]$ respectively. Problem 2: For a test datapoint, the true label and the estimated label are $Y_1 = [0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0]$ and $\hat{Y}_1 = [0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$ respectively.

Table 2 shows the performance of classifiers in these problems obtained using E_{mzo} , E_{ad} , E_{ma}^{cil} , P_{cfci} , E_{rps} , E_{nod}^1 , E_{nod}^2 , and E_{nod}^∞ in example 2. As it can be interpreted from Table 2, E_{mzo} , P_{cfci} , E_{ad} , E_{ma}^{cil} and E_{nod}^∞ treated classifiers of both problems in the same manner. However, the estimated label of the first problem is completely incorrect, while the estimated label in the second problem is very near to the true label. Performance metrics E_{rps} , E_{nod}^1 and E_{nod}^2 reflect the higher performance of the second classifier compared to the first one.

5.3. Example 3

In this example, the main deficiency of E_{rps} and difficulties in interpretation of E_{ad} and E_{ma}^{cil} are exemplified. Consider the following two cases.

Case 1:

For an ordinal three-class classification problem, a completely useless classifier is applied which result in CM_3 .

$$CM_3 = \begin{bmatrix} 0 & 0 & 1 \\ 5000 & 0 & 5000 \\ 0 & 0 & 1 \end{bmatrix} \quad (27)$$

Case 2:

For another ordinal three-class classification problem, con-

Table 2: The performance of two classifiers measured by E_{mzo} , E_{ad} , E_{ma}^{cil} , P_{cfci} , E_{rps} , E_{nod}^1 , E_{nod}^2 , and E_{nod}^∞ in example 2.

Performance Metric	Problem 1	Problem 2
E_{mzo}	1	1
E_{ad}	1	1
E_{ma}^{cil}	1	1
P_{cfci}	0	0
E_{rps}	1	0.0833
E_{nod}^1	1	0.1667
E_{nod}^2	1	0.4082
E_{nod}^∞	1	1

sider a classifier with CM_4 .

$$CM_4 = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 10 & 0 & 0 \end{bmatrix} \quad (28)$$

The performance of classifiers in case 1 and 2 calculated by the E_{mzo} , P_{cfci} , E_{ad} , E_{rps} , E_{nod}^p and E_{ma}^{cil} are listed in Table 3.

As it can be seen from the Table 3, the performance of the applied classifier in case 1 measured by E_{rps} is 0.50, while all estimated labels are incorrect and the classifier is totally useless. The outputs of E_{nod}^p and P_{cfci} are 1 and 0 respectively, which appropriately reflect that the applied classifier is useless in this case. The table also indicates that E_{rps} , E_{ad} and E_{ma}^{cil} result in the same values for both cases, while we know that the applied classifier in the second case is much more effective than the first one. This is appropriately reflected by E_{nod}^1 , E_{nod}^2 and E_{nod}^∞ .

5.4. Example 4

In this example, P_{cfci} , E_{ad} , E_{rps} , and E_{nod}^p are evaluated in measuring the performance of classifiers in C_O^{Pa} . Consider an eight-class C_O^{Pa} . In this problem, the test datapoint label is $Y = [0.1 \ 0.1 \ 0.1 \ 0.1 \ 0.1 \ 0.1 \ 0.2 \ 0.2]$. Two classifiers are applied in this problem. Table 4 shows the output of the applied classifiers. The measured performance of these classifiers using P_{cfci} , E_{ad} , E_{rps} , E_{nod}^1 , E_{nod}^2 and E_{nod}^∞ is presented in Table 5. As it can be understood from Table 4, the estimated label of the second classifier is more similar to the true label compared to that of first classifier. However, the output of the P_{cfci} is the same for both of them. This is due to the fact that the order of categories has no effect on the output of P_{cfci} . In this example, E_{ad} , E_{rps} , E_{nod}^1 , E_{nod}^2 and E_{nod}^∞ reflect the performance improvement from the first classifier to the second one.

Table 3: The performance of two classifiers measured by E_{mzo} , E_{ad} , E_{ma}^{cil} , P_{cfci} , E_{rps} , E_{nod}^1 , E_{nod}^2 , and E_{nod}^∞ in example 3.

Performance Metric	Case 1	Case 2
E_{mzo}	1	0.5
E_{ad}	1	1
E_{ma}^{cil}	1	1
P_{cfci}	0	50
E_{rps}	0.50	0.50
E_{nod}^1	1	0.5714
E_{nod}^2	1	0.5395
E_{nod}^∞	1	0.5

6. Conclusion

In this paper, the ordinal distance between two arbitrary vectors in Euclidean space has been introduced. Then, Normalized Ordinal Distance (E_{nod}^p) as an application-independent performance metric for ordinal, probabilistic-ordinal or partial-ordinal classification problems has been presented. Different advantages of the E_{nod}^p over conventional performance metrics such as mean absolute error of consecutive integer labels E_{ma}^{cil} , mean zero-one error (E_{mzo}), correctly fuzzy classified instances (P_{cfci}), average deviation (E_{ad}), or ranked probability score (E_{rps}) have been shown using a number of numerical examples.

7. Acknowledgements

This work is supported by the European Commission as a Marie-Curie ITN-project (FP7-PEOPLE-ITN-2008), namely Bayesian Biometrics for Forensics (BBfor2), under Grant Agreement number 238803

8. References

- [1] S. Erdural, "A method for robust design of products or processes with categorical response," *METU, Ankara*, 2006.
- [2] M. H. Bahari and H. Van hamme, "Speaker age estimation and gender detection based on supervised non-negative matrix factorization," in *Proc. IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*, 2011, pp. 1–6.
- [3] M.H. Bahari and H. Van hamme, "Speaker age estimation using hidden markov model weight supervectors," in *Proc. 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, 2012, pp. 517–521.
- [4] J.S. Cardoso and J.F.P. da Costa, "Learning to classify ordinal data: the data replication method," *Journal of Machine Learning Research*, vol. 8, no. 1393–1429, pp. 6, 2007.
- [5] J. Verwaeren, W. Waegeman, and B. De Baets, "Learning partial ordinal class memberships with kernel-based proportional odds models," *Computational Statistics and Data Analysis*, vol. 56, no. 4, pp. 928–942, 2012.

Table 4: The output of applied classifiers in example 4.

	Classifier Output (\hat{Y})							
Classifier 1	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1
Classifier 2	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.1

Table 5: The performance of two classifiers measured by P_{cfci} , E_{ad} , E_{rps} , E_{nod}^1 , E_{nod}^2 and E_{nod}^∞ in example 4.

Performance Metric	Classifier 1	Classifier 2
E_{ad}	1.2	0.2
P_{cfci}	80	80
E_{rps}	0.0314	0.0029
E_{nod}^1	0.2927	0.0488
E_{nod}^2	0.2828	0.0853
E_{nod}^∞	0.2222	0.1111

- [6] P. McCullagh, “Regression models for ordinal data,” *Journal of the royal statistical society. Series B (Methodological)*, pp. 109–142, 1980.
- [7] W. Chu and S.S. Keerthi, “Support vector ordinal regression,” *Neural Computation*, vol. 19, no. 3, pp. 792–815, 2007.
- [8] J. Cheng, Z. Wang, and G. Pollastri, “A neural network approach to ordinal regression,” in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, 2008, pp. 1279–1284.
- [9] W. Chu and Z. Ghahramani, “Gaussian processes for ordinal regression,” *Journal of Machine Learning Research*, vol. 6, pp. 1019–1041, 2004.
- [10] SK Shevade and W. Chu, “Minimum enclosing spheres formulations for support vector ordinal regression,” in *Data Mining, 2006. ICDM’06. Sixth International Conference on*. IEEE, 2006, pp. 1054–1058.
- [11] S. Manel, H.C. Williams, and S.J. Ormerod, “Evaluating presence-absence models in ecology: the need to account for prevalence,” *Journal of Applied Ecology*, vol. 38, no. 5, pp. 921–931, 2002.
- [12] E. Van Broekhoven, V. Adriaenssens, and B. De Baets, “Interpretability-preserving genetic optimization of linguistic terms in fuzzy models for fuzzy ordered classification: An ecological case study,” *International Journal of Approximate Reasoning*, vol. 44, no. 1, pp. 65–90, 2007.
- [13] A.M. Mouton, B. De Baets, E. Van Broekhoven, and P.L.M. Goethals, “Prevalence-adjusted optimisation of fuzzy models for species distribution,” *Ecological Modelling*, vol. 220, no. 15, pp. 1776–1786, 2009.
- [14] P. Bougeault, “The wgne survey of verification methods for numerical prediction of weather elements and severe weather events,” *Toulouse: Météo-France*, 2003.
- [15] A.H. Murphy, “On the ranked probability score,” *J. Applied Meteorology*, vol. 8, pp. 988–989, 1969.
- [16] J. Kohonen and J. Suomela, “Lessons learned in the challenge: making predictions and scoring them,” *Lecture Notes in Artificial Intelligence*, pp. 95–116, 2005.
- [17] M. Toda, “Measurement of subjective probability distributions,” Tech. Rep., DTIC Document, 1963.
- [18] M.M. Deza and E. Deza, *Encyclopedia of distances*, Springer, 2009.
- [19] J. Morovic, J. Shaw, and P.L. Sun, “A fast, non-iterative and exact histogram matching algorithm,” *Pattern Recognition Letters*, vol. 23, no. 1, pp. 127–135, 2002.
- [20] J. Luxenburger, *Modeling and Exploiting User Search Behavior for Information Retrieval*, Ph.D. thesis, PhD thesis, Universität des Saarlandes, 2008.
- [21] D.R. Kincaid and E.W. Cheney, *Numerical analysis: mathematics of scientific computing*, vol. 2, Amer Mathematical Society, 2002.
- [22] M. Déqué, JF Royer, R. Stroe, and M. France, “Formulation of gaussian probability forecasts based on model extended-range integrations,” *Tellus A*, vol. 46, no. 1, pp. 52–65, 1994.